

# **A Hybrid Approach to Natural Language Machine Translation for Sinhala and English**



Gray H.P

University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
8/10005  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of MSc in Artificial Intelligence.

**October 2010**

# Declaration

I declare that this dissertation does not incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organization.

Gray H P

Name of Student

Signature of Student



University of Moratuwa, Sri Lanka  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

Date

Supervised by

Prof. A S Karunananda

Name of Supervisor

Signature of Supervisor

Date

# Dedication

I dedicate this thesis to my beloved parents



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Acknowledgements

Firstly, I would like to thank my supervisor Prof. A. S. Karunananda for his help, support, guidance, and many hours of tireless effort throughout the duration of my MSc project. He was a supervisor who showed much effort and enthusiasm. I am grateful to him also for initiating the masters program in artificial intelligence in the University of Moratuwa for the first time in Sri Lanka and giving us this opportunity. He showed much enthusiasm throughout the course of studies and was very punctual for all classes. In fact he showed more enthusiasm than most of us students.

I am also very grateful to all lecturers who were in charge of various subjects and from whom I gathered immense knowledge during the MSc course which is helpful for my career and my project. It is invidious to single people out, but a special thank you to Dr Ruwan Weerasinghe for teaching me Natural Language Processing which was helpful in the project.

I would also like to thank my colleagues at the Moratuwa University. They have been a great source of help with their sharing of knowledge and expertise throughout the MSc duration.

I am also thankful to my employer at Science Land Pvt. Ltd. Mr. A. R. Manamudali for granting me leave from work when it was required to study for my exams and also for providing a part of the fund required for my studies.

I take this opportunity to thank my elder brother Hans Gray for the support he provided and for providing me tips to aid the preparation of this thesis. He also gave me permission to read his DPhil thesis where I gathered much knowledge.

I am extremely thankful to my parents for the time they invested in me and the values they held and instilled in me many years ago. I take this opportunity to thank them for their guidance and encouragement which has helped me not only during my MSc, but more so during my prior education which paved the way to study for my BSc Engineering in University of Peradeniya and MSc in the University of Moratuwa.

## Abstract

Machine Translation is one of the least achieved areas in the area of natural language processing. This is because natural languages are complex, a word can have several meanings, a sentence can have several translations and the translation of a sentence may depend on the context. In this report we describe an approach to machine translation for Sinhala and English languages.

We postulate that humans are able to translate natural languages through simple rules and experience collected without being knowledgeable about sophisticated language construction such as morphology, syntax, semantics and pragmatic structures. This hypothesis has been inspired by the fact that humans construct word forms, phrases and sentences with new words they learn by using simple rules without even being fully conscious about the rules. We do not ignore the fact that all words in a vocabulary do not follow the same rules for forming words. Humans use specific knowledge about certain words when they construct sentences. Also the word selection in a translated sentence varies depending on the context or the semantics of the sentence. Due to this complexity, we focus on a hybrid approach which uses both rules and statistics.

The system described in this thesis focuses on modeling the steps taken by a human to translate a sentence from one language to the other. A bilingual dictionary is used to modal the knowledge of words and synonyms in both languages. Exceptional word dictionaries are used as equivalents to the knowledge of the special words which do not follow the common rules of morphology. The language parsers handle the syntax of sentences in either language. Morphology analyzers are used to handle the rules used in constructing word forms while statistical analyzers are used to handle the proper word usage depending on the syntax.

The system was evaluated by comparing human translation with the machine translation output. The two dominating factors considered were, how understandable the translated sentence is and how much information the translated sentence retains compared to the original. The results are up to the expected quality and further work is required to improve the semantics of translation.

# Contents

Chapter 1	Introduction	1
1.1	Background and Motivation	1
1.2	Aim and Objectives	2
1.3	Solution	2
1.4	Structure	3
Chapter 2	Natural Language Translation – State of the Art	4
2.1	Introduction	4
2.2	Statistical machine translation	4
2.3	Rule based Machine Translation	6
2.4	Summary	6
Chapter 3	The Hybrid Approach	7
3.1	Introduction	7
3.2	The reason for using the hybrid technology	7
3.3	Summary	8
Chapter 4	Hybrid Approach to Natural Language Translation for Sinhala and English	9
4.1	Introduction	9
4.2	Hypothesis	9
4.3	Adopting the Hybrid Technology for Translation	12
4.4	Summary	12
Chapter 5	Design	14
5.1	Introduction	14
5.2	Rule Based Translation	14
5.2.1	Parsers	16
5.2.2	Morphology Analyzers	18
5.2.3	Bilingual Dictionary	19
5.2.4	Parse Tree Translator	20
5.2.5	Transliteration	21
5.3	Summary	23

Chapter 6	Implementation	24
6.1	Introduction	24
6.2	Rule Based Translation	24
6.2.1	Parsers	27
6.2.1.1	English Grammar Rules	27
6.2.1.2	Sinhala Grammar Rules	28
6.2.2	Morphology Analyzers	30
6.2.3	Bilingual Dictionary	33
6.2.4	Parse Tree Translator	34
6.2.5	Transliteration	35
6.3	Development Environment	38
6.4	Summary	39
Chapter 7	Evaluation	40
7.1	Introduction	40
7.2	Human Evaluation (ALPAC)	40
7.3	Results	40
7.4	Testing the Software	41
7.4.1	Translate	42
7.4.2	Find Parse Tree	42
7.4.3	Translate Tree	43
7.4.4	Translate Phrase	44
7.4.5	Find Word	44
7.5	Summary	45
Chapter 8	Conclusion & Further Work	46
8.1	Introduction	46
8.2	Conclusion	46
8.3	Further Work	47
8.4	Summary	47

References	48
Appendix A    Parsers	49
A.1    Code Segments for English Parser	49
A.1.1    Identifying nouns	49
A.1.2    Identifying adjectives	49
A.1.3    Identifying compound adjectives	49
A.1.4    Identifying prepositions	49
A.1.5    Identifying noun phrases	49
A.1.6    Identifying verbs	50
A.2    Code Segments for Sinhala Parser	52
A.2.1    Identifying nouns	52
A.2.2    Identifying adjectives	52
A.2.3    Identifying compound adjectives	52
A.2.4    Identifying verbs	52
A.3    Tenses	54
Appendix B    Morphology	55
B.1    Code Segments for English Morphology Analyzer	55
B.1.1    Changing noun to accusative form	55
B.1.2    Changing a verb to simple present tense	55
B.1.3    Changing a verb to simple past tense	56
B.1.4    Changing a verb to past participle tense	56
B.2    Code Segments for Sinhala Morphology Analyzer	56
B.2.1    Changing noun to accusative form	56
B.2.2    Changing noun to indeterminate form	58
B.2.3    Changing noun to “from” form	58
B.2.4    Changing adjective to “most” form	59
B.3    Examples of Morphology in Sinhala	59
B.3.1    Nouns (noun, pronoun – first, second, third person)	59



B.3.2	Verbs	60
B.3.3	Adjective	61
B.3.4	Adverb	61
B.4	Examples of Morphology in English	61
B.4.1	Nouns (noun, pronoun – first, second, third person)	61
B.4.2	Verbs	63
B.4.3	Adjective	63
B.4.4	Adverb	64
B.5	Common Morphology Rules	64
B.5.1	Nouns	64
B.5.2	Verbs	68
Appendix C	Testing and Quality Assurance	71
C.1	Testing and Quality Assurance	71
C.1.1	Unit testing	71
C.1.2	Integration testing	71
C.1.3	Stress test	71
Appendix D	Schedule	72
D.1	Deliverables	72
D.2	Plan of Action	72

## List of Tables

Table 4-1 : Forming words with an Imaginary noun	9
Table 4-2 : Forming words with an imaginary verb	10
Table 5-1: Bilingual dictionary entities.	20
Table 6-1: ASCII Mapping for Sinhala Characters.	26
Table 6-2: Notations used for grammar rules.	27
Table 6-3: Mappings for English characters.	35
Table 6-4: Mappings for Sinhala characters.	36
Table 6-5: Statistics of syllables in Sri Lankan names.	37
Table 6-6: Examples for issue of the gender.	37
Table 6-7: Examples where gender correction cannot be used.	37
Table 6-8: Initials used in names	38
Table 7-1: Results of translating simple sentences	41
Table A-1: The tenses in language.	54
Table B-1: Morphology of nouns in Sinhala	60
Table B-2: Morphology of verbs in Sinhala	60
Table B-3: Morphology of adjectives in Sinhala	61
Table B-4: Morphology of adverbs in Sinhala	61
Table B-5: Morphology of nouns in English	62
Table B-6: Morphology of verbs in English	63
Table B-7: Morphology of adjectives in English	63
Table B-8: Morphology of adverbs in English	64
Table B-9: Accusative forms in English	64
Table B-10: Accusative forms in Sinhala	65
Table B-11: Plural form in English	65
Table B-12: Plural form in Sinhala	66
Table B-13: Indeterminate form in Sinhala	66

Table B-14: To preposition in Sinhala	66
Table B-15: From preposition in Sinhala	67
Table B-16: Of preposition in Sinhala	67
Table B-17: Simple present tense in English	68
Table B-18: Simple present tense in Sinhala	68
Table B-19: Present continuous tense in English	68
Table B-20: Present continuous tense in Sinhala	69
Table B-21: Simple past tense in English	69
Table B-22: Simple past tense in Sinhala	69
Table B-23: Past continuous tense in English	70
Table D-1: The deliverables of the project	72
Table D-2: Plan of action.	74



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## List of Figures

Figure 4-1: Input Process Output	12
Figure 5-1: Modules in Translator	14
Figure 5-2: Modules in Rule Based Translation	15
Figure 5-3: Changing the state of a word.	18
Figure 5-4: Example for state changing	18
Figure 5-5: Rule based Transliteration	22
Figure 6-1: Flow chart of rule based transliteration.	25
Figure 6-2: Changing the state of a word.	32
Figure 7-1: The testing application	42
Figure 7-2: Evaluating the parse tree of a noun phrase	43
Figure 7-3: Evaluating the translated parse tree of a verb phrase	43
Figure 7-4: Translating a phrase	44
Figure 7-5: Testing the morphology module	45